Computational inference of replication and transcription activator regulator activity in herpesvirus from gene expression data

Recchia, A.; Wit, E.; Vinciotti, V.; Kellam, P.

Published in:
Iet systems biology

DOI:
10.1049/iet-syb:20070053

Published in IET Systems Biology
Received on 9th October 2007
Revised on 10th April 2008
doi: 10.1049/iet-syb:20070053

Special Issue – Selected papers from the 'Omics: Assembling Systems Biology' Workshop

ISSN 1751-8849

Computational inference of replication and transcription activator regulator activity in herpesvirus from gene expression data

A. Recchia[1] *E. Wit*[1] *V. Vinciotti*[2] *P. Kellam*[3]

[1]*Institute of Mathematics and Computing Science, University of Groningen (RUG), 9747 AG Groningen, Netherlands*
[2]*Department of Mathematical Sciences, Brunel University, Uxbridge UB8 3PH, UK*
[3]*Division of Infection and Immunity, University College London, London W1T 4JF, UK*
E-mail: e.c.wit@rug.nl

Abstract: One of the main aims of system biology is to understand the structure and dynamics of genomic systems. A computational approach, facilitated by new technologies for high-throughput quantitative experimental data, is put forward to investigate the regulatory system of dynamic interaction among genes in Kaposi's sarcoma-associated herpesvirus network after induction of lytic replication. A reconstruction of transcription factor activity and gene-regulatory kinetics using data from a time-course microarray experiment is proposed. The computational approach uses nonlinear differential equations. In particular, the quantitative Michaelis–Menten model of gene-regulatory kinetics is extended to allow for post-transcriptional modifications and synergic interactions between target genes and the Rta transcription factor. The kinetic method is developed within a Bayesian inferential framework using Markov chain Monte Carlo. The profile of the Rta transcriptional regulator, other post-transcriptional regulatory genes and gene-specific kinetic parameters are inferred from the gene expression data of the target genes. The method described here provides an example of a principled approach to handle a wide range of transcriptional network architectures and regulatory activation mechanisms to reconstruct the activity of several transcription factors and activation kinetic parameters in a single regulatory network.

1 Introduction

A key aim of system biology is to understand the nature of intra- and intercellular dynamics and to develop mathematical models that describe quantitatively the functional activity and biological interactions among a large number of genes, proteins and other small molecules [1, 2]. By combining experimental and computational approaches, such as high-throughput microarray data and ordinary differential equations (ODEs), we can begin to understand the structure and dynamics of the complex biological systems [3].

Wolkenhauer [4] distinguishes the natural and formal systems, whereby a formal system is an abstract and mathematised representation of the natural system, that is, reality [5]. A common model representation of a dynamical system is via ODEs [6, 7]. One particular use of such ODEs models has been to simulate complex biochemical pathways or pathways with the aim to describe and explain the structure of such biological systems and the resulting data from it. Embedding such models in a statistical inference context has two additional advantages. First, it provides a systematic and principled mechanism for reverse engineering such systems from data. Second, it allows the system itself and, more importantly, the measurements of that system to be subject to random fluctuations, which could make deterministic reconstruction methods inherently unstable.

In recent years, there have been interesting developments in work on inferring transcriptional regulatory networks using ODEs from expression data. Barenco *et al.* [8] used a linear ODE model to find potential targets of a particular transcription factor. Nachman *et al.* [9] proposed a probabilistic method for reconstructing and inferring dynamical model of gene transcription. They used a regulation function to describe the quantitative transcription

rates, assuming a form of Michaelis–Menten (MM) kinetics in which the transcription factor levels were known. Khanin *et al.* [10, 11] developed a statistical framework to reconstruct regulatory activity using an extended MM kinetic model with unknown transcription factor levels. Their model has been implemented for the case of a single-input network motif, consisting of several target genes regulated by a single transcription factor. The kinetic parameters of the gene regulation model were estimated by maximising the likelihood. Izumiya [12] used the same kinetic MM approach of Khanin *et al.* [10, 11] in a Bayesian inference framework.

In contrast to [8], our model uses nonlinear ODEs, allowing for possible saturation effects at high levels of the transcription factor. Our kinetic model allows a very general shape of the transcription factor levels, thereby extending the work by Nachman *et al.* [9]. At the same time, it extends the previous work of Khanin *et al.* [10–12] by considering not only the single-input motifs, but more complicated network structures. Indeed, our model suits network motifs consisting of target genes regulated by one or more transcription factors, which recruits other proteins to form a co-regulating transcriptional process. This model not only infers direct reactions between transcription factor and target genes but also between co-regulating proteins and the activation process itself. In other words, this approach reflects and tests specific biological properties of complex interaction kinetics.

In this article, we construct and analyse a dynamic model that captures the kinetics of a main genetic module activated in the lytic infection phase of Kaposi's sarcoma-associated herpesvirus (KSHV). This module is centred around the Rta transcription factor. We describe the reconstruction of transcription factor activity and gene regulation kinetics. We use data from a designed time-course microarray experiment to quantify mRNA levels and to analyse the herpesvirus gene expression patterns. Our aim is (i) to formalise, (ii) to test and (iii) to extend the existing biological knowledge of the Rta module, embedding a collection of nonlinear ODEs in a statistical framework. The computational approach uses Gibbs sampling and is implemented in a standard and freely available computer package.

2 Rta pathway in lytic infection

The KSHV, also called human herpesvirus eight (HHV-8), was first discovered in 1994 [13] from a Kaposi's sarcoma lesion of an AIDS patient. KSHV is a member of lymphotropic gamma-herpesvirus subfamily with homology to Epstein–Barr virus and Herpesvirus saimiri. Infection by KSHV is a causative agent of three proliferative disorders: Kaposi's sarcoma (KS), primary effusion lymphoma and multicentric Castelman's disease.

The KSHV genome has a double-stranded DNA form of ~165 kb organised into at least 90 open reading frames (ORFs) [14]. Genes in the KSHV genome encode structural proteins that are required for the replication and assembly of new virions. KSHV has two distinct phases in its life cycle, to wit, a latent phase and a lytic replication phase, as shown in Fig. 1. They are characterised by distinct gene expression programmes. During latency only a small subset of viral genes, referred as latent genes, is expressed in an ordered cascade to prevent the death of the infected cell and to maintain the viral genome in a circular, episomal state. Once the virus is reactivated from latency and enters the lytic cycle, many viral genes are transcribed, leading to the production of progeny virions [15]. However, despite its importance for viral propagation and pathogenicity, the nature of the switch from latent infection to lytic replication is still unclear.

Based on their transcriptional kinetics, KSHV genes can be categorised into four more or less distinct classes of genes that were expressed during the lytic replication cycle. These are latent, immediate early, early and late genes [16, 17]. The immediately early genes are expressed after primary infection or reactivation in the initial stage of lytic replication in the presence of inhibitors of protein synthesis. These genes encode regulatory proteins that activate the cascade of early gene expression essential for viral DNA replication and for the regulation of late gene expression. In [18] the lytic genes are temporally similarly classified as primary (activated within 0–10 h after lytic infection), secondary (10–24 h) and tertiary (48–72 h), based on gene expression pattern by DNA arrays. This classification suggests that genes with the same function tend to have similar expression profile.

The switch from latency to lytic KSHV infection is initiated by the Rta. Rta is an immediately early viral transcript product of the open reading frame 50 (ORF50) that affects the expression of viral and host cellular genes. Ectopic overexpression of the Rta protein alone is both necessary and sufficient to disrupt viral latency and to induce a lytic reaction [19–23].

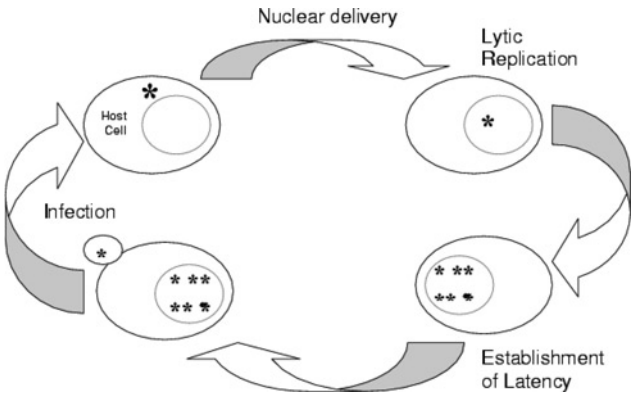

Figure 1 *Simplified life cycle of KSHV*

(i) The virus attaches the host cell surface; (ii) inside the host cell the DNA viruses are uncoated and transported to the host nucleus; (iii) the establishment of the latent phase; (iv) the start of the lytic phase because of the expression of viral genes leading to the synthesis of virus proteins and (v) the virions are transported from nucleus outside the host cell in order to infect a new cell

This phase of lytic replication leads to the death of the host cell

doi: 10.1049/iet-syb:20070053

Numerous studies suggest that Rta transactivates the cellular and viral promoters of a multitude of lytic genes, which have been implicated in virus replication.

In this article, we examine the Rta protein activation of a set of target lytic virus genes, namely K8 (KbZIP), ORF57 (Mta), polyadenylated nuclear (PAN) RNA (nut-1, T1.1, K7), Kaposin (K12), viral interleukin-6 (vIL-6 or K2), K5, K9 (vIRF1), ORF59 (PF8), thymidine kinase (TK or ORF21), viral G protein-coupled receptor (vGPCR or ORF74), K14, ORF6 (single-stranded DNA binding protein) and its own transcript ORF50 [20, 24–28]. We also investigate the role of some regulatory proteins that modify the Rta-induced expression of these target genes. Fig. 2 shows the schematic overview of the regulatory Rta pathway in KSHV. Oval shapes represent proteins and rectangles represent RNA transcripts. All target genes are activated by the transcription factor, Rta, indicated by an arrow. Arrows from RNA transcripts to their associated proteins represent the natural translation process, whereas arrows from secondary proteins, KbZIP, ORF57 and ORF74, to RNA transcripts represent transcription modification by these proteins of the RNA transcription process. The literature suggests that KbZIP down-modulates its own expression and expression of ORF57 [29–31], ORF74 protein down-modulates Rta expression [32], and ORF57 protein up-modulates ORF59 and PAN expression [33].

3 Modelling the Rta pathway

The aims of this paper are to encapsulate the existing biological knowledge, described in the previous section, into a mathematical model and via time-course transcriptional data to test the accuracy of this knowledge and to give it a quantitative context.

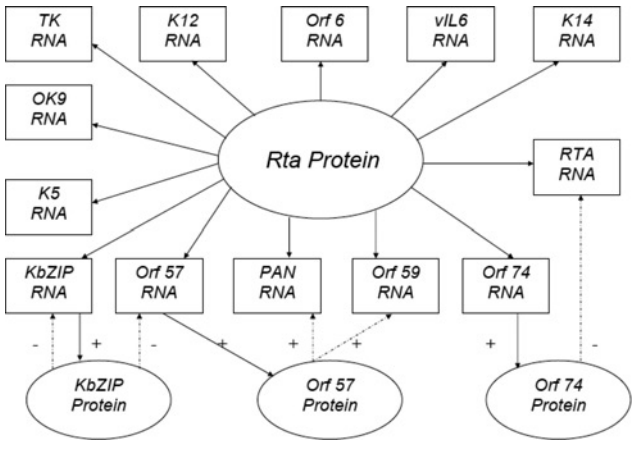

Figure 2 *Schematic representation of the Rta pathway in KSHV*

Rta transcription factor activates 13 target genes – for two of these targets, ORF57 and ORF74, the schematic shows their protein translation

The dashed lines represent the transcription modification by KbZIP, ORF57, ORF74 proteins of some of the target gene activations by Rta

3.1 Microarray time-course experiment

As part of this study, Dr. P. Kellam's Lab (University College, London) performed a microarray experiment, in which totally 68 dual-channel cDNA microarrays were hybridised. The hybridised samples consisted of a human cell-line 293T (clone 6) infected with KSHV and was sampled at 30 different time points after the induction of lytic infection with a mixture of 12-o-tetradecanoyl phorbol-13-acetate (TPA) (20 ng/ml) and Butyrate (3 mM) in 0.5 ml medium (10% foetal calf serum (FCS), P/S, 6 μl/ml hygromycin).

Thirty time points were sampled, equally spaced between 0 and 58 h after the start of the lytic infection. Each time point was sampled twice (biological replicates) and each sample was hybridised twice (technical replicates). Furthermore, 16 reference hybridisation samples were prepared and inserted into an interwoven loop design [34, 35] at regular intervals. Together this provided the original 136 hybridisation samples required.

After the experiment, 10 microarrays were omitted from further analysis, because of low cell numbers and loss of samples during extraction. Among those 10 arrays were all the four arrays that contained samples obtained at 6 h after lytic infection. The remaining 116 channels were subjected to data cleansing methodologies, such as spatial correction, dye normalisation and a form of global quantile normalisation described in [36], using the `smida` R package [37]. The resulting data were used for the analysis described in the following sections.

3.2 Differential equation model for mean expression levels

Typically, microarray experiments destructively sample their hybridisation samples. Therefore the time-course data from the experiment described above cannot be interpreted as a usual time-series data, in which correlations between nearby observations play a crucial role. In fact, in principle, all of our observations are completely independent, except for dependence induced by using technical replicates and the bivariate structure of dual-channel microarrays. For this reason, we extend the nonlinear ODE framework introduced by Khanin *et al.* [10] to model the mean kinetics of these time-course data.

As is well-known from general enzyme kinetics, the average rate of change in expression of a regulated gene is described by the number of RNA molecules 'produced' by the transcription factor and the amount of RNA decay per unit of time. In other words, a simple model for the average kinetic expression $\mu_i(t)$ of gene i at time t is described via the following differential equation

$$\dot{\mu}_i(t) = p_i(t, \eta) - \delta_i \mu_i(t) \quad (1)$$

in which $p_i(t, \eta)$ represents an RNA production term or transcription rate of gene i in the presence of an amount η of its transcription factor, δ_i the linear RNA decay rate of gene i and $\dot{\mu}_i(t)$ represents the derivative of $\mu_i(t)$. The production term depends on the activity of transcription factor η and gene-specific kinetic parameters. We propose to use an extension of the MM kinetics

$$p_i(t, \eta) = \alpha_i + \beta_i \frac{\eta(t)}{\gamma_i + \eta(t)} \tag{2}$$

where α_i, β_i, and γ_i are the gene kinetic parameters of the target gene i. The additive constant α_i accounts for the basal level of transcription in the absence of the regulator and possibly for nuisance background effects from microarray, which have been unaccounted for. The value β_i is the maximum induction in the rate of production of gene i by the transcription factor and γ_i is the half-saturation constant for gene i, reflecting how efficiently and quickly the transcription factor is able to induce the production of RNA of gene i. The general solution of an ODE given by (1) and (2) is given as

$$\mu_i(t) = \left(\mu_{0i} - \frac{\alpha_i}{\delta_i}\right) e^{-\delta_i t} + \frac{\alpha_i}{\delta_i} + \int_0^t \beta_i e^{-\delta_i(t-s)} \frac{\eta(t)}{\gamma_i + \eta(t)} ds$$

where μ_{0i} is an arbitrary constant. This model is used for all target genes without synergistic interactions to K12, ORF6, TK, K9, ORF74, K14, vIL6 and K5, which are all activated via the unobserved Rta transcription factor levels, η_{Rta}. By allowing η to have a quite general functional form, there is typically no explicit solution to the integral, but simple numeric methods can be used to solve it in practice.

For target genes with synergistic relationships with other modulating proteins, we extend the MM model by allowing the maximum amount of transcription, β_i, to depend on additional regulators. We propose a parsimonious functional form of β_i for a given gene i and a given modulating protein η_m by

$$\beta_i(t) = \max\{\psi_{0,i} + \psi_{1,i} \times \eta_{\text{m}}(t), 0\}$$

which is linear in the amount of protein of the gene that modulates the activation by the transcription factor Rta. It is only sensible that β_i cannot become negative and therefore we impose saturation when β_i hits 0. In this model, the sign of the parameter $\psi_{1,i}$ has a natural interpretation of whether or not the modulating protein is a down-modulator or up-modulator of Rta activation. Table 1 summarises the equation terms used in the proposed model.

In our method, we model explicitly the activation of Rta transcription by the Rta transcription factor, but leave out the translational relationship between the two, which anyway is poorly identifiable without any strong modelling assumptions due to the absence of any data of the Rta protein.

Table 1 List of equation terms

Kinetic parameters	Meaning of parameters
$\mu_i(t)$	expression due to mRNA production of gene i at time t
$\dot{\mu}_i(t)$	change in expression of $\mu_i(t)$, as kinetic expression of gene i at time t
$\eta(t)$	level of the transcription factor
$p_i(t, \eta)$	RNA rate of transcription production of gene i by a protein η
δ_i	rate of linear mRNA degradation of gene i
α_i	basal level of transcription of gene i
β_i	maximum rate of production for gene i
γ_i	half-saturation constant for gene i
$\mu_{0,i}$	arbitral constant for gene i
$\psi_{0,i}$	expression per hour for gene i
$\psi_{1,i}$	expression per protein per hour for gene i
$y_i(t_{\text{Cy3}})$	data for gene i on channel Cy3
$y_i(t_{\text{Cy5}})$	data for gene i on channel Cy5
$\mu_i(t_{\text{Cy3}})$	deterministic gene expression for gene i on channel Cy3
$\mu_i(t_{\text{Cy5}})$	deterministic gene expression for gene i on channel Cy5
ε_{S}	random spot effect
ε_{T}	random technical variation for gene i on channel Cy3
ε'_{T}	random technical variation for gene i on channel *Cy*5

3.3 Statistical model

Given a model for the mean expression levels of all the genes involved in the Rta pathway, the next step is to relate the actual observations of the designed experiment to this model. It is generally accepted that log-transformed microarray data have their variances approximately stabilised [36]. Moreover, the fact that dual-channel microarray data come in pairs, as a result of the natural channel pairing on each array, means that we have to account for possible correlation between the Cy3 and Cy5 channel data on the arrays. Therefore we propose here to model the data via a nonlinear mixed effects model. For each of the 13 target genes i of the Rta pathway,

doi: 10.1049/iet-syb:20070053

we model the logarithm of the two channels for a particular microarray as

$$\log y_i(t_{\mathrm{Cy3}}) = \log \mu_i(t_{\mathrm{Cy3}}) + \epsilon_{\mathrm{S}} + \epsilon_{\mathrm{T}}$$
$$\log y_i(t_{\mathrm{Cy5}}) = \log \mu_i(t_{\mathrm{Cy5}}) + \epsilon_{\mathrm{S}} + \epsilon'_{\mathrm{T}}$$

where t_j is the condition, that is, time point or reference, in channel j, $\mu_i(t)$ the deterministic gene expression model from Section 3.2, $\epsilon_{\mathrm{S}} \sim N(0, \sigma_{\mathrm{S}}^2)$ a random spot effect and $\epsilon_{\mathrm{T}} \sim N(0, \sigma_{\mathrm{T}}^2)$ random technical variation.

3.4 Inference

As done previously in [12], we use a Bayesian inferential framework to estimate the profile of Rta transcriptional regulator, other regulatory protein levels and gene-specific kinetic parameters of our kinetic model. The inference code was written in the WinBUGS programming language, a freely available software [38].

We define the parameter vector, $\boldsymbol{\vartheta}_i = (\alpha_i, \beta_i, \gamma_i, \delta_i, \mu_{0,i}, \psi_{0,i}, \psi_{1,i})$, for each gene i $(i = 1, \ldots, p)$ and the overall parameter vector of the model,

$$\theta = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_p, \eta_{\mathrm{Rta}}, \eta_{\mathrm{KbZIP}}, \eta_{\mathrm{ORF57}}, \eta_{\mathrm{ORF74}}, \sigma_{\mathrm{T}}, \sigma_{\mathrm{S}})$$

The parameters have a joint posterior probability density $p(\theta \mid y) \propto p(y \mid \theta)p(\theta)$, where the likelihood expression is given by

$$p(y|\theta) = \prod_{j=1}^{n} p\big((y(t_{j,\mathrm{Cy3}}), y(t_{j,\mathrm{Cy5}}))|\theta\big)$$
$$= \prod_{j=1}^{n} p(y(t_{j,\mathrm{Cy3}})|\theta)p(y(t_{j,\mathrm{Cy3}})|y(t_{j,\mathrm{Cy5}}), \theta)$$

and where we assume a simple jointly independent prior distribution for θ. We take inverse gamma priors on the variance parameters σ_{S} and σ_{T}, exponential priors on the gene-specific kinetic parameters, α_i, γ_i, δ_i and μ_{0i}. We

Table 2 Posterior estimates of kinetic parameters α and β

Parameter	Mean	SD	MC error	2.5%	97.5%
basal level of production					
α_{OK9}	2.85	2.44	0.12	0.10	9.26
α_{K5}	3.37	2.84	0.12	0.13	10.78
α_{K12}	7.55	5.47	0.28	0.44	21.25
α_{ORF6}	3.00	2.78	0.12	0.10	10.46
α_{ORF21}	3.90	3.46	0.16	0.12	12.81
α_{K14}	0.99	1.01	0.02	0.03	3.88
α_{K2}	1.00	1.00	0.02	0.02	3.63
α_{ORF57}	2.04	1.39	0.06	0.09	5.37
α_{ORF74}	4.16	3.91	0.19	0.12	14.91
α_{ORF59}	1.01	1.01	0.02	0.03	3.80
α_{KbZIP}	1.77	1.41	0.05	0.07	5.28
α_{PAN}	1.01	0.99	0.02	0.03	3.72
α_{Rta}	1.06	1.04	0.02	0.03	3.84
maximum rate of production					
β_{OK9}	17.68	2.70	0.09	12.27	22.81
β_{K5}	20.13	3.12	0.09	13.76	25.85
β_{K12}	18.49	5.08	0.22	6.59	27.33
β_{ORF6}	21.75	3.27	0.09	15.48	28.44
β_{ORF21}	16.96	3.41	0.11	9.80	23.03
β_{K14}	0.95	0.97	0.02	0.03	3.65
β_{K2}	1.00	0.97	0.02	0.03	3.75

Table 3 Posterior-estimates of kinetic parameters δ, γ, ψ_1

Parameter	Mean	SD	MC error	2.5%	97.5%
rate of linear mRNA degradation					
δ_{OK9}	0.02	0.01	0.00	0.01	0.04
δ_{K5}	0.02	0.01	0.00	0.01	0.04
δ_{K12}	0.02	0.01	0.00	0.01	0.05
δ_{ORF6}	0.02	0.01	0.00	0.01	0.04
δ_{ORF21}	0.02	0.01	0.00	0.01	0.04
δ_{K14}	1.01	1.00	0.02	0.03	3.69
δ_{K2}	1.04	1.02	0.02	0.04	3.64
δ_{ORF57}	0.01	0.00	0.00	0.01	0.02
δ_{ORF74}	0.03	0.01	0.00	0.01	0.06
δ_{ORF59}	2.18	1.17	0.09	0.65	5.28
δ_{KbZIP}	0.01	0.00	0.00	0.01	0.03
δ_{PAN}	1.92	1.12	0.05	0.48	4.82
δ_{Rta}	2.21	1.08	0.06	0.74	4.96
half saturation constant					
γ_{OK9}	0.54	0.33	0.01	0.14	1.42
γ_{K5}	0.52	0.30	0.01	0.14	1.28
γ_{K12}	0.45	0.34	0.01	0.08	1.35
γ_{ORF6}	0.51	0.30	0.01	0.14	1.29
γ_{ORF21}	0.52	0.35	0.01	0.11	1.37
γ_{K14}	0.98	0.98	0.02	0.02	3.71
γ_{K2}	0.99	1.03	0.02	0.02	3.75
γ_{ORF57}	0.63	0.57	0.02	0.08	2.12
γ_{ORF74}	0.51	0.31	0.01	0.12	1.33
γ_{ORF59}	0.01	0.01	0.00	0.00	0.02
γ_{KbZIP}	0.56	0.46	0.01	0.06	1.81
γ_{PAN}	0.02	0.02	0.00	0.00	0.06
γ_{Rta}	0.05	0.02	0.00	0.01	0.09
expression per protein per hour					
$\psi_{1,ORF57}$	5.65	1.43	0.06	3.29	8.79
$\psi_{1,ORF59}$	2.12	0.90	0.07	0.72	4.21
$\psi_{1,KbZIP}$	7.47	1.59	0.06	4.66	11.02
$\psi_{1,PAN}$	1.79	0.88	0.04	0.51	3.80
$\psi_{1,Rta}$	3.23	1.27	0.07	1.25	6.02

IET Syst. Biol., 2008, Vol. 2, No. 6, pp. 385–396
doi: 10.1049/iet-syb:20070053

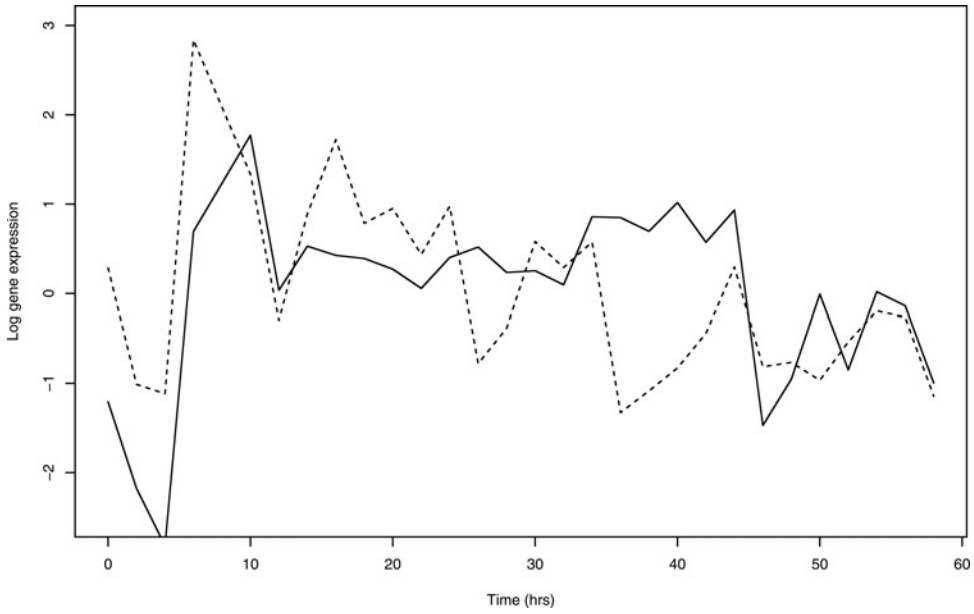

Figure 3 *Observed expression profile of the regulated genes PAN (dashed line) and Rta (solid line) in response to lytic induction by the Rta transcription factor*

place an exponential prior on β_i of target genes without synergic relationship and activated only by Rta transcription factor, normal priors on the parameters ψ_0 and ψ_1 for genes, which are synergically regulated by Rta and a modifying protein.

The WinBUGS package implements a clever Gibbs sampling algorithm to generate samples from the joint posterior distribution [39]. We run two separate chains with different initial values. The chains convergence speed seems to be affected by the initial values, but after 50 000 iterations, the Markov Chain Monte Carlo (MCMC) sampler in both chains seems to be sufficient to reach the convergence.

4 Results

Tables 2 and 3 report the posterior means, posterior standard deviations, with 95% posterior credible intervals, of the MM kinetic parameters for each gene of Rta pathway. The MC error in these tables indicates the error associated with the random nature of the MCMC, which decreases as the number of MCMC samples increases. In all cases, the MC

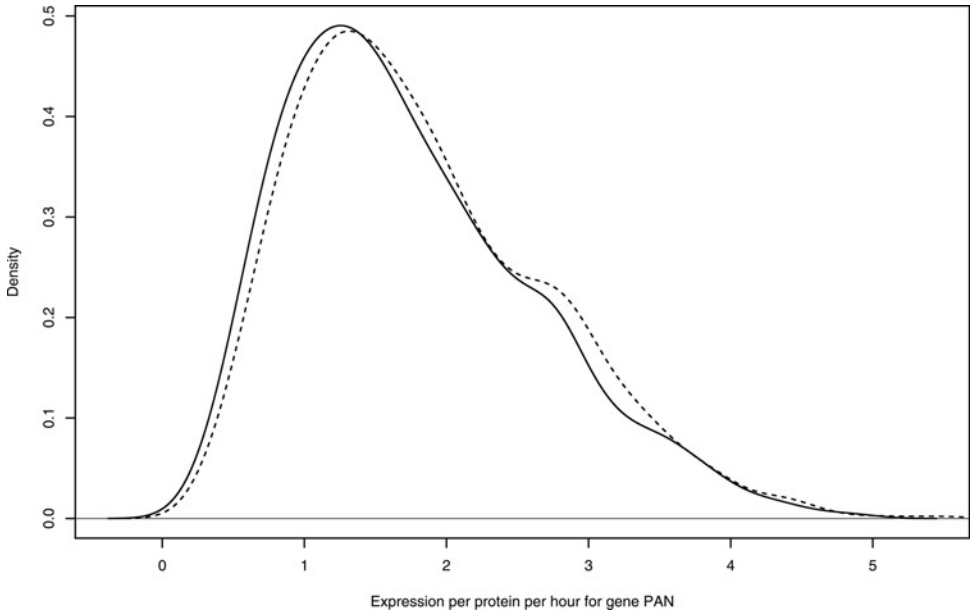

Figure 4 *$\psi_{1,PAN}$ posterior density for PAN gene for two MCMC chains*

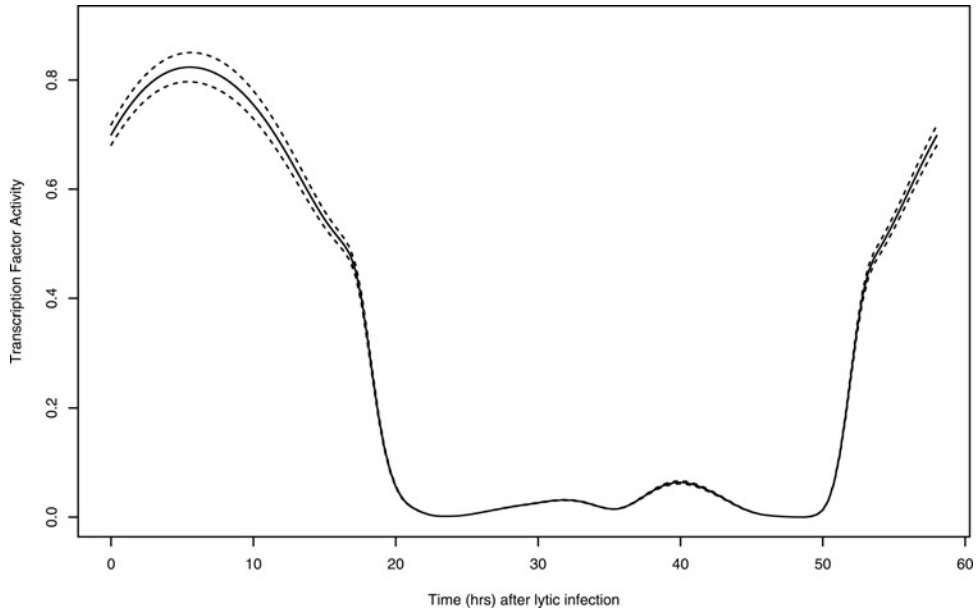

Figure 5 *Reconstructed Rta transcription factor profile (solid line)*
The dashed lines are the 2.5 and 97.5 percentiles of the posterior

errors seem ignorable. Parameter names are related to the model in Section 3.2.

From the results, it is clear that K2 and K14 do not seem to be regulated by Rta, as was predicted in the literature. The level of β is extremely small, effectively consistent with no Rta activation whatsoever. In fact, it seems that there is very little other information about these genes in the data, as the estimates for most of its kinetic parameters are consistent with the original prior.

The degradation rates δ for several genes (OK9, K5, K12, ORF6, ORF21, ORF57, ORF74 and KbZIP) are quite small ($\delta \simeq 0.01$), which given the overall absolute expression levels of about 500, almost cancel out their basal production rates. Other genes (PAN, ORF59 and ORF50) degrade much faster ($\delta \simeq 2$), which makes their degradation an integral part of their expression profile.

The half-saturation constants, γ, suggest that all target genes, except K14 and K2, are immediately early or early genes, supporting the kinetic gene classification during the lytic cycle. From our results, we can roughly make a distinction between two groups of genes. We find that ORF59, Rta and PAN respond very quickly to changes in Rta ($\gamma \simeq 0.01$), whereas the other eight genes have significantly slower response rates ($\gamma \simeq 0.5$). These results are in qualitative agreement with those obtained in [17, 21, 31], except that [40, 41], respectively, classify PAN and ORF59 as early, rather than an immediate-early genes, such as Rta.

The immediate reaction of the target genes at induction into lytic replication, generated by Rta, is shown in Fig. 3. The Rta transcription factor rapidly induces the up-regulation of PAN expression and its own gene Rta. The dashed line represents the PAN expression and the solid line the transcription factor expression after lytic infection. During the first hours, both genes show a rapid decrease of their expression, with a peak level observed between 2 and 3 h, following an immediate increase in maximum expression at 4 and 5 h. The expression of both genes is high in 4–20 h and during the last hours decrease.

The basal level of production, α, for the target genes may indicate whether the expression of some genes might be influenced by some additional transcription factors other than Rta. We obtained relatively high values for genes K12, ORF74, K9 and K5 ($\alpha \simeq 5$), which may indicate an additional regulator. However, the posterior standard errors of these estimates are quite large and therefore there is very little evidence to support this in one way or another.

The parameters ψ_1 capture the synergistic relationship between the transcription factor Rta and several of its potential modulators. Fig. 4 shows the posterior density estimate of $\psi_{1,\mathrm{PAN}}$ for the two MCMC chains. Both estimates $\hat{\psi}_{1,\mathrm{ORF59}} = 2.1(\mathrm{SD}0.9)$ and $\hat{\psi}_{1,\mathrm{PAN}} = 1.8(\mathrm{SD}0.9)$ suggest a positive synergic relationship between Rta and ORF57, which is confirmed in the existing literature [33]. However, the estimates $\hat{\psi}_{1,\mathrm{Rta}} = 3.2(\mathrm{SD}1.3)$, the effect of ORF74 on the transcription of Rta [32, 42], $\hat{\psi}_{1,\mathrm{KbZIP}} = 7.5(\mathrm{SD}\ 1.6)$, the effect of KbZIP on its own transcription [30, 31] and $\hat{\psi}_{1,\mathrm{ORF57}} = 5.6(\mathrm{SD}\ 1.4)$, the effect of KbZIP on the transcription of ORF57 [29], are all positive, in contrast to the quoted literature, which suggest that genes are down-regulated, respectively, by ORF74 and KbZIP.

doi: 10.1049/iet-syb:20070053

The reconstructed profile by MCMC approach of the main transcription factor of the target genes is shown in Fig. 5. The reconstructed Rta profile (Fig. 5) shows a periodic trend, properly reflecting the periodic nature of Rta profile (Fig. 3). The solid line represents the inferred mean profile, whereas the dashed lines are the 2.5 and 97.5 percentiles. The reconstructed Rta profile shows high activity at the very beginning of the lytic infection until about 10 h afterwards, which then drops down to zero by 18 h. This is followed by a period, in which the transcription factor activity is essentially zero, until around 50 h post-infection it rises again. Although some of the target genes show up-regulation of expression at the end of the experimental period (e.g. Fig. 6), there is no obvious biological reason for this late flare of Rta activity. It is hypothesised that perhaps a secondary lytic infection wave has been the cause of this. Comparing the transcription factor reconstruction in Fig. 5 and its associated transcription profile in Fig. 3, it is clear that the Rta activity anticipates its expression by ~2 h.

Fig. 6 compares the transcription data relative to the reconstructed expression profiles for four genes, PAN, ORF57, ORF59 and KbZIP. In general, we notice that the inferred profiles for these genes show a good fit with the observed expression data, keeping in mind the non-trivial correlation structure in the data. It may seem that there is a lot of variation in the data, but it should be remembered that there are various variance components in the data, each of which can be estimated from the data.

There is some correlation between the parameters, summarised in Fig. 7. It shows the correlation matrix for all the parameters in the model in a schematic fashion. Roughly, black indicates negative, grey positive and white no or low correlations. The *x*- and *y*-axes represent the parameters ordered in an alphabetic fashion. The strongly correlated parameters are explicitly indicated on the graph. The high values along the diagonal suggest, not surprisingly, that there is a lot of internal (time) autocorrelation for the parameters

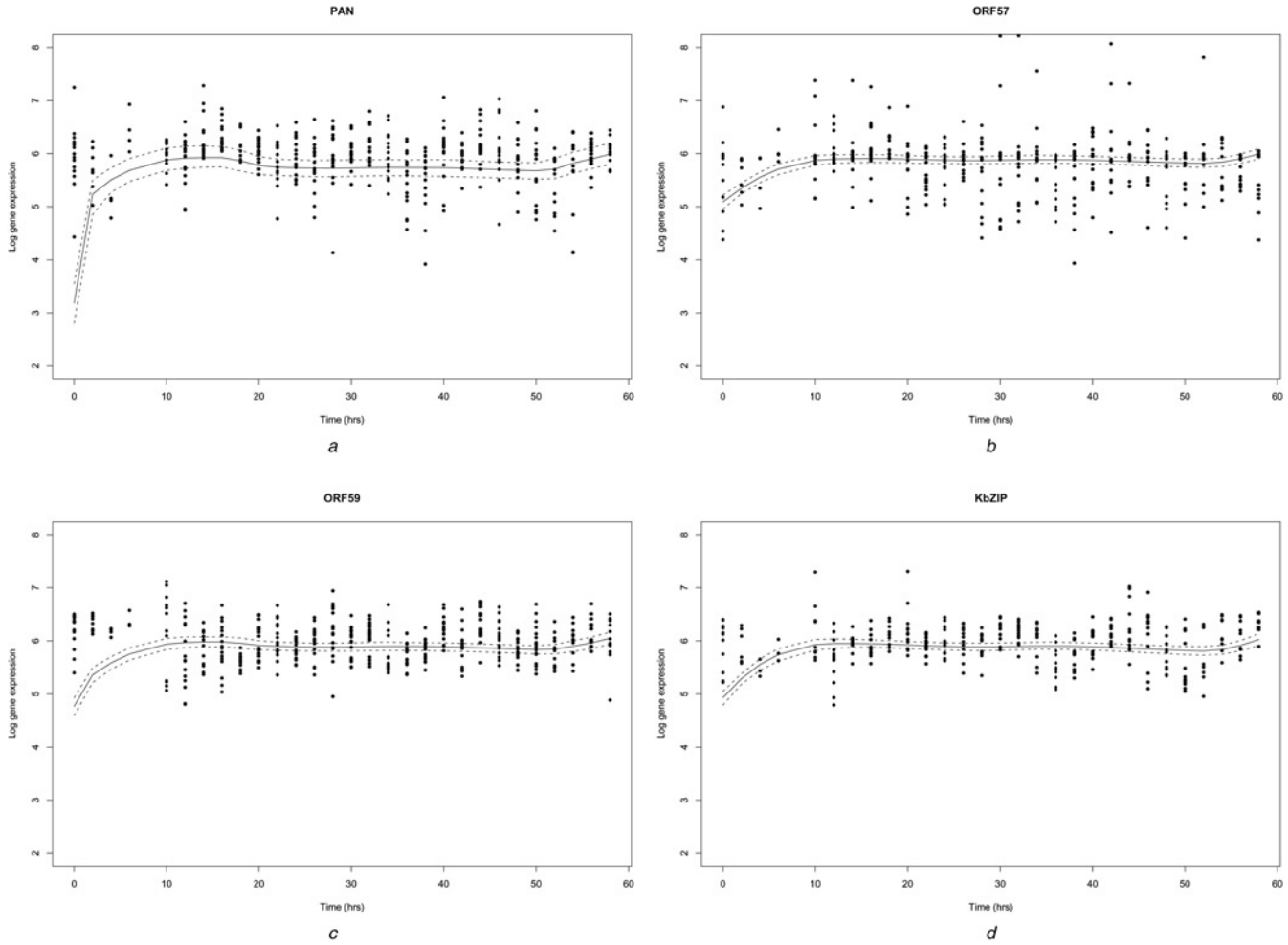

Figure 6 *Microarray data and reconstructed expression profiles for four target genes*

a PAN
b ORF57
c ORF59
d KbZIP
The marks are the observed data for four technical replicates generated for each gene, the solid lines stand for the estimated gene profile

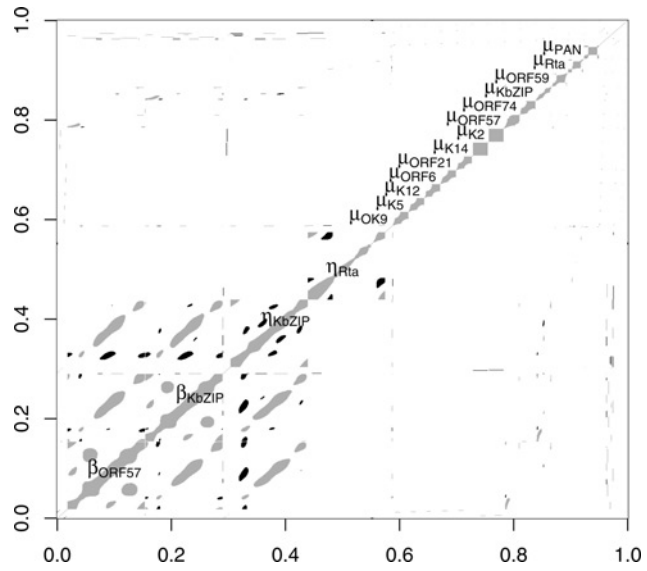

Figure 7 *Schematic correlation matrix for the posterior of the parameters in the model (black: Cor* ≤ -0.5*; white:* $-0.5 <$ *Cor* ≤ 0.5*; grey: Cor* ≥ -0.5*)*

β_{ORF57}, β_{ORF74}, η_{KbZIP}, η_{Rta} and the various average expression profile reconstructions μ. Slightly anomalously, μ_{K2} and μ_{K14} are almost totally autocorrelated, due to the fact that these profiles are particularly flat. Correlation is quite strong between β_{ORF57} and β_{KbZIP}, reflecting the fact that they both are functions of η_{KbZIP}. The sporadic cases of negative autocorrelation within parameter posteriors η_{Rta} and η_{KbZIP} are the result of the periodic behaviour of the profiles as can be seen in Fig. 5.

5 Conclusions

In this paper, we have set out a computational approach for beginning to understand a complex biological system by means of a combination of (i) carefully sifting through the biological literature to identify general information about the structure of the system and the nature of its dynamics, (ii) an appropriate mathematical model to capture this dynamics and (iii) a statistical model to capture the sampling scheme and the variability of the data.

Building forth on the existing MM kinetic models, we have introduced a novel extension for reconstructing, quantifying and inferring the transcriptional regulatory network in KSHV. Previously published statistical models for describing regulatory activity levels, either tended to use simple linear ODEs [8] or restricted their attention on simple network motifs [9–12]. Our method does not sacrifice network complexity for statistical tractability and as such is an example of what can be achieved with routinely available genomic data. The method uses data from two-channel microarrays in a nonlinear mixed effects model, incorporating a mixture of fixed effects, such as the average gene expression profiles, and random effects, such as spot effects and technical variations.

The main aim of this study was to examine the central role of the Rta transcription factor in the initiation of lytic KSHV infection. We found that Rta triggers the lytic infection cycle and that it quickly activates the transcription of Rta, PAN and ORF59, before, more slowly, activating the transcription of another eight genes. Contrary to what was known from the literature, two of the genes, namely K14 and K2, do not seem to be activated by Rta at all. We have shown that Rta expression induces the up-regulation of its target genes, besides being involved in synergetic relationships with KbZIP, ORF74 and ORF57 in the transactivation of several target genes. Some of these results correspond to the existing literature on the matter, whereas others are novel and need to be tested further. In particular, from the posterior estimations, we have that: the two genes K2 and K14 are immediately early genes and are not regulated by Rta transcription factor [18]; ORF59, Rta and PAN have significantly higher response rates [17, 21, 31]; the positive synergic relationship between Rta and ORF57 [33] and KbZIP protein has an effect on its transcription and on ORF57 [29–31]; this in contrast to the repression of KbZIP and ORF74 [29, 32, 42].

Our study also provides novel estimates for degradation rates, which are not available in the literature. Results indicate that some genes, such as OK9, K5, K12, ORF6, ORF21, ORF57, ORF74 and KbZIP, degrade slowly; others, such as PAN, ORF59 and ORF50, degrade much faster. Therefore we believe that our model, although initially set out to model the existing biological information, adds useful quantitative information. The statistical approach proposed in this paper provides a principled approach to handle microarray data for a wide range of transcriptional network architectures and regulation functions to reconstruct the activity of several transcription factors in a complex regulatory network.

6 References

[1] KITANO H.: 'Foundations of system biology' (MIT Press, Cambridge, 2001)

[2] WOLKENHAUER O., KITANO H., CHO K.H.: 'An introduction to systems biology: opportunities and challenges for physical scientists in the postgenomic ere of the biomedical sciences', *IEEE Control Syst.*, 2003, **24**, (4), pp. 38–48

[3] KITANO H.: 'Computational systems biology', *Nature*, 2002, **420**, pp. 206–210

[4] WOLKENHAUER O.: 'Mathematical modelling in the post-genome era: understanding genome expression and regulation – a system theoretic approach', *Biosystem*, 2002, **65**, pp. 1–18

[5] WOLKENHAUER O., MESAROVIC M.: 'Feedback dynamics and cell function: Why system biology is called system biology?', *Mol. Biosyst.*, 2005, **1**, pp. 14–16

doi: 10.1049/iet-syb:20070053

[6] CHEN P., HE H.L., CHURCH G.M.: 'Modeling gene expression with differential equations'. Proc. Pacific Symp. Biocomputing, 1999, vol. 4, pp. 29–40

[7] SMOLEM P., BAXTER D.A., BYRNE J.H.: 'Modeling transcriptional control in gene networks', *Bull. Math. Biol.*, 2000, **62**, (2), pp. 247–292

[8] BARENCO M., TOMESCU D., BREWER D., CALLARD R., STARK J., HUBANK M.: 'Ranked prediction of p53 targets using hidden variable dynamic modeling', *Genome Biol.*, 2006, **7**, (3), p. R25

[9] NACHMAN I., REGEV A., FRIEDMAN N.: 'Infering quantitative models of regulatory networks from expression data', *Bioinformatics*, 2004, **20**, (1), pp. i248–i256

[10] KHANIN R., VINCIOTTI V., WIT E.: 'Reconstructing repressor protein levels from expression of gene targets in *Escherichia coli*', *PANS*, 2006, **103**, (49), pp. 18592–18596

[11] KHANIN R., VINCIOTTI V., MERSINIAS M., SMITH C., WIT E.: 'Statistical reconstruction of transcription factor activity using Michaelis–Menten kinetics', *Biometrics*, 2007, **63**, (3), pp. 816–823

[12] ROGERS S., KHANIN R., GIROLAMI M.: 'Bayesian model-based inference of transcription factor activity', *BMC Bioinf.*, 2007, **8**, (Suppl 2), p. S2

[13] CHANG Y., CESARMAN E., PESSIN M.S., *ET AL.*: 'Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma', *Science*, 1994, **266**, (5192), pp. 1865–1869

[14] NEIPEL F., ALBRECHT J.C., FLECKENSTEIN B.: 'Cell-homologous genes in the Kaposi's sarcoma-associated rhadinovirus human herpesvirus 8: determinants of its pathogenicity', *J. Virol.*, 1997, **71**, pp. 4187–4192

[15] RENNE R., ZHONG W., HERNDIER B., *ET AL.*: 'Lytic growth of Kaposi's sarcoma-associated herpesvirus (human herpesvirus 8) in culture', *Nat. Med.*, 1996, **2**, pp. 342–346

[16] WEST J.T., WOOD C.: 'The role of Kaposi's sarcoma-associated herpesvirus/human herpesvirus-8 regulator of transcription activation (Rta) in control of gene expression', *Oncogene*, 1993, **22**, pp. 5150–5163

[17] SUN R., LIN S.F., STASKUS K., *ET AL.*: 'Kinetics of Kaposi's sarcoma-associated herpesvirus gene expression', *J. Virol.*, 1999, **73**, pp. 2232–2242

[18] JENNER R.G., ALBA M.M., BOSHOFF C., KELLAM P.: 'Kaposi's sarcomaassociated herpesvirus latent and lytic gene expression as revealed by DNA arrays', *J. Virol.*, 2001, **75**, pp. 891–902

[19] GRADOVILLE L., GERLACH J., GROGAN E., *ET AL.*: 'Kaposi's sarcoma-associated herpesvirus open reading frame 50 Rta protein activates the entire viral lytic cycle in HH-B2 primary effusion lymphoma cell line', *J. Virol.*, 2000, **74**, pp. 6207–6212

[20] LUKAC D.M., KIRSHNER J.R., GANEM D.: 'Transcriptional activation by the product of open reading frame 50 of Kaposi's sarcoma-associated herpesvirus is required for lytic viral reactivation in B cells', *J. Virol.*, 1999, **73**, pp. 9348–9361

[21] SUN R., LIN S.F., GRADOVILLE L., YUAN Y., ZHUAND F., MILLER G.: 'A viral gene that activates lytic cycle expression of Kaposi's sarcoma-associated herpesvirus', *PNAS*, 1999, **95**, (18), pp. 10866–10871

[22] NAKAMURA H., LU M., GWACK Y., SOUVLIS J., ZEICHNER S.L., JUNG J.U.: 'Global changes in Kaposi's sarcoma-associated virus gene expression patterns following expression of a tetracycline-inducible Rta transactivator', *J. Virol.*, 2003, **77**, pp. 4205–4220

[23] SONG M.J., DENG H., SUN R.: 'Comparative study of regulation of Rta-responsive genes in Kaposi's sarcoma-associated herpesvirus/human herpesvirus 8', *J. Virol.*, 2003, **77**, (17), pp. 9451–9462

[24] DENG H., SONG M.J., CHU J.T., SUN R.: 'Transcriptional regulation of the interleukin-6 gene of human herpesvirus 8 (Kaposi's sarcoma-associated herpesvirus)', *J. Virol.*, 2002, **76**, pp. 8252–8264

[25] LUKAC D.M., RENNE R., KIRSHNER J.R., GANEM D.: 'Reactivation of Kaposi's sarcoma-associated herpesvirus infection from latency by expression of the ORF 50 transactivator, a homolog of the EBV R protein', *Virology*, 1998, **252**, (2), pp. 304–312

[26] KRISHNAN H.H., NARANATT P.P., SMITH M.S., ZENG L., BLOOMER C., CHANDRAN B.: 'Concurrent expression of latent and a limited number of lytic genes with immune modulation and antiapoptotic function by Kaposi's sarcom-aassociated herpesvirus early during infection of primary endothelial and fibroblast cells and subsequent decline of lytic gene expression', *J. Virol.*, 2004, **78**, (7), pp. 3601–3620

[27] BOWSER B.S., MORRIS S., SONG M.J., DAMANIA B.: 'Characterization of Kaposi's sarcoma-associated herpesvirus (KSHV) k1 promoter activation by rta', *Virol.*, 2006, **348**, (2), pp. 309–327

[28] SAKAKIBARA S., UEDA K., CHEN J., OKUNO T., YAMANISHI K.: 'Octamerbinding sequence is a key element for the autoregulation of Kaposi's sarcoma-associated herpesvirus ORF50/lyta gene expression', *J. Virol.*, 2001, **75**, pp. 6894–6900

[29] IZUMIYA Y., LIN S.F., ELLISON T., *ET AL.*: 'Kaposi's sarcoma-associated herpesvirus K-BZIP is a coregulator of K-Rta: physical association and promoter-dependent transcriptional repression', *J. Virol.*, 2003, **77**, pp. 1441–1451

[30] IZUMIYA Y., ELLISON T., YEH E.T.H., *ET AL.*: 'Kaposi's sarcoma-associated herpesvirus K-BZIP represses gene transcription via SUMO modification', *J. Virol.*, 2005, **79**, pp. 9912–9925

[31] LIAO W., TANG Y., LIN S.F., KUNG H.J., GIAM C.Z.: 'K-BZIP of Kaposi's sarcoma-associated herpesvirus/human herpesvirus 8 (KSHV/HHV-8) binds KSHV/HHV-8 Rta and represses Rta-mediated transactivation', *J. Virol.*, 2003, **77**, (6), pp. 3809–3815

[32] CANNON M.L., CESARMAN E.: 'KSHV G protein-coupled receptor inhibits lytic gene transcription in primary-effusion lymphoma cells via p21-mediated inhibition of cdk2', *Blood*, 2006, **107**, (1), pp. 277–284

[33] KIRSHNER J.R., LUKAC D.M., CHANG J., GANEM D.: 'Kaposi's sarcoma-associated herpesvirus open reading frame 57 encodes a posttranscriptional regulator with multiple distinct activities', *J. Virol.*, 2000, **74**, (8), pp. 3586–3597

[34] WIT E., NOBILE A., KHANIN R.: 'Near-optimal designs for dual-channel microarray studies', *JRSS-C*, 2005, **54**, (5), pp. 817–830

[35] VINCIOTTI V., KHANIN R., D'ALIMONTE D., *ET AL.*: 'An experimental evaluation of a loop versus a reference designs for two-channel microarrays', *Bioinformatics*, 2005, **21**, pp. 492–501

[36] WIT E., MCCLURE J.: 'Statistics for microarray' (Wiley, New York, 2004)

[37] smida R package, http://www.stats.gla.ac.uk/microarray/book

[38] MRC Biostatistics Unit Cambridge, http://www.mrc-bsu.cam.ac.uk/bugs

[39] GELMAN A., CARLIN J.B., STERN H.S., RUBIN D.B.: 'Bayesian data analysis' (Chapman & Hall, 2004)

[40] SONG M.J., BROWN H.J., WU T.T., SUN R.: 'Transcription activation of polyadenylated nuclear RNA by Rta in human herpesvirus 8/Kaposi's sarcoma-associated herpesvirus', *J. Virol.*, 2001, **75**, (7), pp. 3129–3140

[41] ZHU F.X., CUSANO T., YUAN Y.: 'Identification of the immediate-early transcripts of Kaposi's sarcoma-associated herpesvirus', *J. Virol.*, 1999, **73**, (7), pp. 5556–5567

[42] CANNON M.L., PHILPOTT N.J., CESARMAN E.: 'The Kaposi's sarcoma-associated herpesvirus G protein-coupled receptor has broad signaling effects in primary effusion lymphoma cells', *J. Virol.*, 2003, **77**, pp. 57–67

doi: 10.1049/iet-syb:20070053